

Pipeline Complet de Modélisation Mathématique et Optimisation des Modèles pour le Diagnostic du COVID-19 et l'Analyse de son Spectre Clinique

Abdeljalil Idalahaj

December 6, 2025

Contents

1 Notations Générales	3
2 Prétraitement des Données	3
2.1 Imputation des valeurs manquantes	3
2.2 Encodage des variables catégorielles	3
2.3 Normalisation et transformation	4
3 Feature Engineering et Sélection	4
4 Modèles Mathématiques	4
4.1 Régression Logistique	4
4.2 Gradient Boosting / LightGBM	5
4.3 SVM — Support Vector Machine	5
4.3.1 Problème primal (soft margin)	5
4.3.2 Lagrangien	5
4.3.3 Forme duale	5
4.3.4 Noyaux (Kernel trick)	5
4.3.5 Fonction de décision	6
4.3.6 Conditions KKT	6
4.3.7 Optimisation SMO	6
4.4 Modèle de Cox (Survie)	6
5 Calibration et Probabilités	6
6 Optimisation des Hyperparamètres — GridSearchCV	7
7 Décision et Seuil Optimal	7
8 Interprétabilité — SHAP	7
9 Approche Bayésienne	7

10 Validation Statistique et Tests Rigoureux	8
10.1 Cross-validation et Nested CV	8
10.2 Comparaisons Rigoureuses	8
11 Méthodes de Robustesse et Fairness	8
11.1 Sensibilité aux Missingness	8
11.2 Analyse par Sous-Groupes	8
12 Formules Utiles et Recettes de Calcul	8
12.1 Log-loss (binaire)	8
12.2 Brier Score	9
12.3 AUC empirique (Mann-Whitney)	9
12.4 Bootstrap pour IC	9
12.5 Gain d'un split (Boosting)	9

1 Notations Générales

Données :

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n, \quad x_i \in \mathbb{R}^d$$

Cibles :

$$y_i \in \{0, 1\} \text{ (binaire)}, \quad y_i \in \{1, \dots, K\} \text{ (multi-classe)}$$

Données de survie :

$$(t_i, \delta_i, x_i), \quad \delta_i \in \{0, 1\}$$

Modèles paramétriques :

$$f_\theta(x), \quad s_\theta(x), \quad \hat{p}_\theta(x) = \sigma(s_\theta(x))$$

Vecteur de labels :

$$y = (y_1, \dots, y_n)^\top$$

Matrice de confusion : TP, FP, TN, FN.

2 Prétraitement des Données

2.1 Imputation des valeurs manquantes

Indices des valeurs manquantes :

$$M_j = \{i : x_{ij} \text{ manquant}\}$$

Imputation moyenne :

$$\tilde{x}_{ij} = \bar{x}_j = \frac{1}{n_j} \sum_{i \notin M_j} x_{ij}$$

Imputation par régression :

$$x_{ij} = \beta^\top x_{i,-j} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$\hat{x}_{ij} = \hat{\beta}^\top x_{i,-j}$$

Imputation multiple (Rubin) :

$$\bar{\theta} = \frac{1}{m} \sum_{k=1}^m \hat{\theta}^{(k)}, \quad T = \bar{U} + \left(1 + \frac{1}{m}\right) B$$

2.2 Encodage des variables catégorielles

One-hot :

$$e(C) \in \{0, 1\}^m$$

Target encoding lissé :

$$\tilde{y}_c = \frac{n_c \bar{y}_c + \alpha \bar{y}}{n_c + \alpha}$$

2.3 Normalisation et transformation

Standardisation :

$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

Transformation logarithmique :

$$x' = \log(1 + x)$$

Condition number :

$$\kappa(X^\top X) = \frac{\lambda_{\max}}{\lambda_{\min}}$$

3 Feature Engineering et Sélection

Création de nouvelles features :

$$z_{ab} = x_a x_b, \quad x_j^2, \quad x_j^3$$

LASSO :

$$\hat{\theta} = \arg \min_{\theta} \{L(\theta) + \lambda \|\theta\|_1\}$$

Test t pour différenciation :

$$t = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{s_1^2/n_1 + s_0^2/n_0}}$$

Variance Inflation Factor (VIF) :

$$VIF_j = \frac{1}{1 - R_j^2}$$

4 Modèles Mathématiques

4.1 Régression Logistique

$$\begin{aligned} \log \frac{p_\theta(x)}{1 - p_\theta(x)} &= \theta^\top x \\ p_\theta(x) &= \sigma(\theta^\top x), \quad \sigma(u) = \frac{1}{1 + e^{-u}} \end{aligned}$$

Log-vraisemblance :

$$\ell(\theta) = \sum_i [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$$

Gradient :

$$\nabla L = - \sum_i (y_i - p_i) x_i + 2\lambda\theta$$

Hessien :

$$H = \sum_i p_i(1 - p_i) x_i x_i^\top + 2\lambda I$$

Newton-Raphson :

$$\theta_{t+1} = \theta_t - H^{-1} \nabla L$$

Intervalle de confiance et odds ratio :

$$\hat{\theta}_j \pm 1.96 \sqrt{H_{jj}^{-1}}, \quad OR_j = e^{\hat{\theta}_j}$$

4.2 Gradient Boosting / LightGBM

Initialisation :

$$F_0(x) = \arg \min_{\gamma} \sum_i \ell(y_i, \gamma)$$

Itération :

$$F_m(x) = F_{m-1}(x) + \eta h_m(x)$$

Approximation Taylor de la loss :

$$\ell \approx \ell + g_i h(x_i) + \frac{1}{2} h_i h(x_i)^2$$

Poids de feuille optimal :

$$w_j^* = -\frac{\sum_{i \in R_j} g_i}{\sum_{i \in R_j} h_i + \lambda}$$

Gain :

$$G(R_j) = \frac{1}{2} \frac{(\sum g_i)^2}{\sum h_i + \lambda} - \gamma$$

4.3 SVM — Support Vector Machine

4.3.1 Problème primal (soft margin)

$$J(w, b, \xi) = \frac{1}{2} \|w\|^2 + C \sum_i \xi_i$$

Sous contraintes :

$$y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

4.3.2 Lagrangien

$$\begin{aligned} \mathcal{L} &= \frac{1}{2} \|w\|^2 + C \sum_i \xi_i - \sum_i \alpha_i (y_i(w^\top x_i + b) - 1 + \xi_i) - \sum_i \mu_i \xi_i \\ \alpha_i &\geq 0, \quad \mu_i \geq 0 \end{aligned}$$

4.3.3 Forme duale

$$W(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

Sous contraintes :

$$0 \leq \alpha_i \leq C, \quad \sum_i \alpha_i y_i = 0$$

4.3.4 Noyaux (Kernel trick)

$$x_i \cdot x_j \longrightarrow K(x_i, x_j)$$

Exemples :

$$K(x, z) = x \cdot z, \quad K(x, z) = (\gamma x \cdot z + r)^d$$

$$K(x, z) = \exp(-\gamma \|x - z\|^2), \quad K(x, z) = \tanh(\gamma x \cdot z + r)$$

4.3.5 Fonction de décision

$$f(x) = \text{sign}\left(\sum_i \alpha_i y_i K(x_i, x) + b\right)$$

$$p(y = 1|x) = \frac{1}{1 + \exp(-Af(x) - B)}$$

4.3.6 Conditions KKT

$$\alpha_i[y_i(w \cdot x_i + b) - 1 + \xi_i] = 0, \quad \mu_i \xi_i = 0$$

4.3.7 Optimisation SMO

$$E_i = f(x_i) - y_i$$

$$\alpha_i^{new} = \alpha_i^{old} + \frac{y_i y_j (E_j - E_i)}{\eta}, \quad \eta = 2K(x_i, x_j) - K(x_i, x_i) - K(x_j, x_j)$$

$$\alpha_i \in [0, C]$$

$$b^{new} = b^{old} - E_i - y_i(\alpha_i^{new} - \alpha_i^{old})K(x_i, x_i) - y_j(\alpha_j^{new} - \alpha_j^{old})K(x_i, x_j)$$

4.4 Modèle de Cox (Survie)

$$h(t|x) = h_0(t)e^{\beta^\top x}$$

Log-vraisemblance partielle :

$$\ell(\beta) = \sum_{i:\delta_i=1} \left[\beta^\top x_i - \log \sum_{j:t_j \geq t_i} e^{\beta^\top x_j} \right]$$

Gradient :

$$\frac{\partial \ell}{\partial \beta} = \sum_{i:\delta_i=1} \left[x_i - \frac{\sum_{j:t_j \geq t_i} x_j e^{\beta^\top x_j}}{\sum_{j:t_j \geq t_i} e^{\beta^\top x_j}} \right]$$

5 Calibration et Probabilités

Platt scaling :

$$\hat{p}_i = \sigma(as_i + b)$$

Brier score :

$$B = \frac{1}{n} \sum_i (\hat{p}_i - y_i)^2$$

AUC :

$$\widehat{AUC} = \frac{1}{n_+ n_-} \sum_{i:y_i=1} \sum_{j:y_j=0} \left(I[s_i > s_j] + \frac{1}{2} I[s_i = s_j] \right)$$

Bootstrap pour IC :

$$IC_{95\%} = [q_{0.025}, q_{0.975}]$$

Tests statistiques : DeLong et McNemar

$$z = \frac{A_1 - A_2}{\sqrt{Var(A_1 - A_2)}}, \quad \chi^2 = \frac{(b - c)^2}{b + c}$$

6 Optimisation des Hyperparamètres — GridSearchCV

Fonction objectif :

$$Score(\theta) = \frac{1}{k} \sum_{r=1}^k Score_{fold_r}(\theta)$$

Optimisation :

$$\theta^* = \arg \max_{\theta} Score_{CV}(\theta)$$

Exemples par modèle :

$$\min \left(\frac{1}{2} \|w\|^2 + C \sum \xi_i \right) \text{ (SVM)}$$

$$\min (-L(\theta) + \lambda \|\theta\|^2) \text{ (logistique)}$$

$$\min \sum_i \ell(y_i, F_i(x_i)) \text{ (LightGBM)}$$

Validation croisée :

$$\hat{m}_{CV} = \frac{1}{k} \sum_{r=1}^k m^{(r)}, \quad Var(\hat{m}) = \frac{1}{k-1} \sum_{r=1}^k (m^{(r)} - \hat{m})^2$$

7 Décision et Seuil Optimal

Seuil de décision optimal :

$$\tau^* = \arg \min_{\tau} [C_{FN} FN(\tau) + C_{FP} FP(\tau)]$$

$$d^*(x) = 1 \iff \hat{p}(x) \geq \frac{C_{FP}}{C_{FP} + C_{FN}}$$

8 Interprétabilité — SHAP

Valeur de contribution de chaque feature :

$$\phi_j(x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f_{S \cup \{j\}}(x) - f_S(x)]$$

9 Approche Bayésienne

Posterior :

$$p(\theta|D) \propto \prod_i p_\theta(y_i|x_i)p(\theta)$$

10 Validation Statistique et Tests Rigoureux

10.1 Cross-validation et Nested CV

k-fold CV : partitionner le jeu de données en k folds. Pour chaque fold r , entraîner le modèle sur $\mathcal{D} \setminus \mathcal{D}_r$ et évaluer sur \mathcal{D}_r .

Estimation de la métrique :

$$\hat{m}_{CV} = \frac{1}{k} \sum_{r=1}^k m^{(r)}$$

Nested CV : - Outer loop : estimation de la performance réelle du modèle. - Inner loop : sélection des hyperparamètres. Permet d'éviter l'optimisme dans l'évaluation.

10.2 Comparaisons Rigoureuses

Pour comparer deux modèles A et B sur k -folds : - Test de Wilcoxon sur les différences des métriques par fold (non-paramétrique) - Test t si la normalité est plausible

Remarque : Attention aux dépendances entre folds. Il est préférable d'utiliser des répétitions de CV et de réaliser les tests sur les réplications.

11 Méthodes de Robustesse et Fairness

11.1 Sensibilité aux Missingness

Simuler des valeurs manquantes selon MCAR, MAR, MNAR, paramétré par $\pi_j(x)$. Mesurer la décroissance de l'AUC en fonction du taux de missing r :

$$AUC(r)$$

Courbe analysée et modélisée par régression linéaire ou LOESS.

11.2 Analyse par Sous-Groupes

Pour chaque sous-groupe G :

$$m_G, \quad IC \text{ par bootstrap}$$

Comparer les sous-groupes via des tests d'hypothèse, par exemple test de proportions pour la sensibilité.

12 Formules Utiles et Recettes de Calcul

12.1 Log-loss (binaire)

$$LL = -\frac{1}{n} \sum_{i=1}^n \left[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \right]$$

12.2 Brier Score

$$B = \frac{1}{n} \sum_{i=1}^n (\hat{p}_i - y_i)^2$$

12.3 AUC empirique (Mann-Whitney)

$$\widehat{AUC} = \frac{\sum_{i:y_i=1} r_i - \frac{n_+(n_++1)}{2}}{n_+ n_-}$$

où r_i est le rang des scores, n_+ et n_- sont respectivement le nombre de positifs et négatifs.

12.4 Bootstrap pour IC

Tirer B échantillons bootstrap et prendre les quantiles correspondants pour l'intervalle de confiance.

12.5 Gain d'un split (Boosting)

$$Gain = \frac{1}{2} \left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right) - \gamma$$

avec

$$G = \sum_i g_i, \quad H = \sum_i h_i$$